SPACE ALTERNATING VARIATIONAL ESTIMATION AND KRONECKER STRUCTURED DICTIONARY LEARNING

Christo Kurisummoottil Thomas, Dirk Slock

EURECOM, Sophia-Antipolis, France, Email:{kurisumm,slock}@eurecom.fr

ABSTRACT

In this paper, we address the fundamental problem of Sparse Bayesian Learning (SBL), where the received signal is a high-order tensor. We furthermore consider the problem of dictionary learning (DL), where the tensor observations are assumed to be generated from a Kronecker structured (KS) dictionary matrix multiplied by the sparse coefficients. Exploiting the tensorial structure results in a reduction in the number of degrees of freedom in the learning problem, since the dimensions of each of the factor matrices are significantly smaller than the matricized dictionary if we vectorize the observations. We propose a novel fast algorithm called space alternating variational estimation with dictionary learning (SAVED-KS), which is a version of variational Bayes (VB)-SBL pushed to the scalar level. Similarly, as for SAGE (space-alternating generalized expectation maximization) compared to EM, the component-wise approach of SAVED-KS compared to SBL renders it less likely to get stuck in bad local optima and its inherent damping (more cautious progression) also leads to typically faster convergence of the non-convex optimization process. Simulation results show that the proposed algorithm has a faster convergence rate and lower mean squared error (MSE) compared to the alternating least squares (ALS) based method for tensor decomposition.

Index Terms— Sparse Bayesian Learning, Variational Bayes, Tensor Decomposition, Dictionary Learning, Alternating Optimization

1. INTRODUCTION

In many applications such as Multiple Input Multiple Output (MIMO) radar [1], massive MIMO channel estimation [2], image and video processing etc., the received signals are multidimensional (i.e tensors). Moreover, these signals can be represented as a low rank tensor. To fully exploit the structure of such signals, tensor decomposition methods such as CANDECOMP/PARAFAC (CP) [3,4] or Canonical Polyadic Decomposition (CPD) [5] have been introduced. In this paper, we consider a generalized problem in which the dictionary matrix can be factorized as a Kronecker product [6], the received tensor signal $\mathbf{Y}$ can be represented as,

$$\mathbf{y} = (\mathbf{A}_1 \otimes \mathbf{A}_2 .... \otimes \mathbf{A}_N)\mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{y} = vec(\mathbf{Y})$, $\otimes$ represents the Kronecker product between two matrices, $vec(\cdot)$ representing the vectorized version of the tensor or matrix $(\cdot)$, $\mathbf{Y} \in \mathcal{C}^{I_1 \times I_2 \ldots \times I_N}$ is the observations or data, $\mathbf{A}_{j,i} \in \mathcal{C}^{I_j}$, the factor matrix $\mathbf{A}_j = [\mathbf{A}_{j,1}, ..., \mathbf{A}_{j,P_j}]$ which is unknown and the tensor product is represented by $[\![\mathbf{A}_1, ..., \mathbf{A}_N; \mathbf{x}]\!]$, $\mathbf{x}$ is the $M(= \prod_{j=1}^{N} P_j)$-dimensional sparse signal and $\mathbf{w}$ is the additive noise. $\mathbf{x}$ contains only K non-zero entries, with $K << M$ and thus the dictionary matrix to be learned allows a low rank representation. $\mathbf{w}$ is assumed to be a white Gaussian noise, $\mathbf{w} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$. To address this problem when the dictionary matrix is known, a variety of algorithms such as the orthogonal matching pursuit [7], the basis pursuit method [8] and the iterative re-weighted l_1 and l_2 algorithms [9] exist in the literature. The SBL introduced by [10, 11], is developed around a sparsity-promoting prior for $\mathbf{x}$, whose realizations are softly sparse in a sense that most entries are small in magnitude and close to zero.

CPD can be viewed as a general extension of the singular value decomposition (SVD) to the high-order tensors, with the difference that the factor matrices need not be orthogonal. In certain applications such as wireless channel estimation, these factors have specific forms such as Vandermonde or Toeplitz or Hankel. To find the tensor factor matrices, the most popular solution is the ALS [12], which iteratively optimizes one factor matrix at a time while keeping the others fixed. Most of the existing algorithms [13–17] focus on either maximum likelihood based schemes, LS or K-SVD algorithms. Knowledge of tensor rank is a prerequisite to implement these algorithms and it takes large number of iterations for them to converge. Moreover, classical algorithms ignore the potential statistical knowledge of the factor matrices into account. While we focus on a Bayesian approach to the estimation of the factor matrices in this paper, with automatic relevance determination.

1.1. Contributions of this paper

- We propose a novel Space Alternating Variational Estimation based SBL technique with KS dictionary learning called SAVED-KS, advancing the SAVE methods which we introduced in [18–21] which assumed a known or Khatri-Rao structured dictionary.
- We also propose a joint VB version for the KS dictionary matrix factors which has better performance compared to SAVED-KS, but at the cost of an increase in computational complexity.
- We also discuss the local identifiability using the non-singularity of the Fisher information matrix (FIM) for KS DL in a SBL setting.
- Numerical results suggest that the proposed solution has a faster convergence rate (and hence lower complexity) than (even) the classical ALS and furthermore has lower reconstruction MSE in the presence of noise.

In the following, boldface lower-case and upper-case characters denote vectors and matrices respectively. The operators tr$(\cdot)$, $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$, $\|\cdot\|$ represents trace, transpose, conjugate, conjugate

EURECOM's research is partially supported by its industrial members: ORANGE, BMW, ST Microelectronics, Symantec, SAP, Monaco Telecom, iABG, and by the projects DUPLEX (French ANR) and MASS-START (French FUI).

transpose and Frobenius norm respectively. A complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Theta}$ is distributed as $x \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Theta})$. diag $(\cdot)$ represents the diagonal matrix created by elements of a row or column vector. The operator $< x >$ or $\mathrm{E}(\cdot)$ represents the expectation of x. All the random variables are complex here unless specified otherwise. We represent $\bigotimes_{j=1}^{N} \mathbf{A}_j = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes ... \otimes \mathbf{A}_N$. $\mathbf{I}_N$ represents the identity matrix with dimension N. We define the unfolding operation on an N^{th} order tensor $\mathbf{Y} = [\![\mathbf{A}_1, ..., \mathbf{A}_N; \mathbf{x}]\!]$ as [12] ($\mathbf{Y}^{(n)}$ is of size $I_n \times \prod_{i=1, i\neq n}^{N} I_i$ below),

$$\mathbf{Y}^{(n)} = \mathbf{A}_n \mathbf{X}^{(n)} (\mathbf{A}_N \otimes \mathbf{A}_{N-1} ... \mathbf{A}_{n+1} \otimes \mathbf{A}_{n-1} ... \otimes \mathbf{A}_1)^T. \tag{2}$$

2. HIERARCHICAL PROBABILISTIC MODEL

In the following sections, we represent (3) using the tensor decomposition properties from [12]. Let $Y_{i_1,...,i_N}$ represents the $(i_1, i_2, ..., i_N)^{th}$ element of the tensor and $\mathbf{y} = [y_{1,1,...,1}, y_{1,1,...,2}, ..., y_{i_1,i_2,...,i_N}]^T$, then it can be verified that [22, 23],

$$\mathbf{y} = (\mathbf{A}_1 \otimes \mathbf{A}_2 ... \otimes \mathbf{A}_N)\mathbf{x} + \mathbf{w} = (\bigotimes_{j=1}^{N} \mathbf{A}_j)\mathbf{x} + \mathbf{w}, \tag{3}$$

where we denote $\mathbf{A} = \bigotimes_{j=1}^{N} \mathbf{A}_j$. Since the sparsity measure (number of nonzero components) of $\mathbf{x}$ is unknown and the following VB-SBL algorithm performs automatic rank determination. In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\mathbf{x}$ as in [10]. The hierarchical prior is chosen such that it encourages the sparsity property of $\mathbf{x}$. $\mathbf{x}$ is assumed to have a Gaussian distribution parameterized by $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ ... \ \alpha_M]$, where α_i (which is a real quantity) represents the inverse variance or the precision parameter of x_i, $p(\mathbf{x}/\boldsymbol{\alpha}) = \prod_{i=1}^{M} p(x_i/\alpha_i) = \prod_{i=1}^{M} \mathcal{CN}(0, \alpha_i^{-1})$. Further a Gamma prior is considered over $\boldsymbol{\alpha}$, $p(\boldsymbol{\alpha}) = \prod_{i=1}^{M} p(\alpha_i/a, b) = \prod_{i=1}^{M} \Gamma^{-1}(a) b^a \alpha_i^{a-1} e^{-b\alpha_i}$. The inverse of noise variance γ is also assumed to have a Gamma prior, $p(\gamma) = \Gamma^{-1}(c) d^c \alpha_i^{c-1} e^{-d\gamma}$. Now the likelihood distribution can be written as,

$$p(\mathbf{y}/\mathbf{x}, \mathbf{A}, \gamma) = (2\pi)^{-N} \gamma^N e^{-\gamma ||\mathbf{y} - \mathbf{A}\mathbf{x}||^2}. \tag{4}$$

We emphasize that the presented algorithm do not exploit parametric forms, because those parametric forms are uncertain. For eg., considering the massive MIMO channel estimation problem [24], the array response at the mobile station (MS) is not exploitable. Even the array response at the base station (BS) will typically require calibration to be exploitable. Doppler shifts are clear Vandermonde vectors. Delays could be more or less clear, if one goes to frequency domain in OFDM, and one only takes into account the range of subcarriers for which the Tx/Rx filters can be considered frequency-flat. Then over those subcarriers, it's also Vandermonde. We consider $\mathbf{A}_{j,i} = [1 \ \mathbf{a}_{j,i}^H]^H$ and further $\mathbf{a}_{j,i}$ is unconstrained and deterministic (in all the Vandermonde cases, it is perfect, or in all cases of phasors). Assuming first entry to be 1 is even better than $\|\mathbf{A}_{j,i}\| = 1$ because $\|\mathbf{A}_{j,i}\| = 1$ still leaves a phase ambiguity. With first entry$= 1$, the factors are unique, up to permutation in the sum of terms. It is to be noted that one major difference compared to the DL for Khatri-Rao structured matrix factors as looked upon in our paper [21] is that, we avoid considering a discretized dictionary and instead the sparsity for $\mathbf{x}$ comes from considering the cases of multi-paths with same delay having different AoA or AoDs.

2.1. Variational Bayesian Inference

The computation of the posterior distribution of the parameters is usually intractable. In order to address this issue, in variational Bayesian framework, the posterior distribution $p(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}/\mathbf{y})$ is approximated by a variational distribution $q(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A})$ that has the factorized form:

$$q(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}) = q_\gamma(\gamma) \prod_{i=1}^{M} q_{x_i}(x_i) \prod_{i=1}^{M} q_{\alpha_i}(\alpha_i) \prod_{j=1}^{N} \prod_{i=1}^{P_j} q_{\mathbf{a}_{j,i}}(\mathbf{a}_{j,i}). \tag{5}$$

Variational Bayes compute the factors q by minimizing the Kullback-Leibler distance between the true posterior distribution $p(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}/\mathbf{y})$ and the $q(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A})$. From [25], The KL divergence minimization is equivalent to maximizing the evidence lower bound (ELBO) [26]. Doing this in an alternating fashion for each variable in $\boldsymbol{\theta}$ leads to (a more detailed discussion in our paper [18]),

$$\begin{aligned} \ln(q_i(\theta_i)) &= < \ln p(\mathbf{y}, \boldsymbol{\theta}) >_{k \neq i} + c_i, \\ p(\mathbf{y}, \boldsymbol{\theta}) &= p(\mathbf{y}/\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}) p(\mathbf{x}/\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\gamma), \end{aligned} \tag{6}$$

where $\boldsymbol{\theta} = \{\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}\}$ and θ_i represents each scalar in $\boldsymbol{\theta}$. Here $<>_{k \neq i}$ represents the expectation operator over the distributions q_k for all $k \neq i$.

3. SAVED-KS SPARSE BAYESIAN LEARNING

In this section, we propose a Space Alternating Variational Estimation (SAVE) based alternating optimization between each elements of $\boldsymbol{\theta}$. For SAVE, not any particular structure of $\mathbf{A}$ is assumed, in contrast to AMP which performs poorly when $\mathbf{A}$ is not i.i.d or sub-Gaussian. Based on a quadratic loss function, the Bayesian estimator of a parameter is the posterior mean; we therefore define the variational Bayesian estimators of parameters $\boldsymbol{\theta}$ as the means of the variational approximation to the posterior distribution. The joint distribution can be written as,

$$\begin{aligned} &\ln p(\mathbf{y}, \boldsymbol{\theta}) = N \ln \gamma - \gamma \, ||\mathbf{y} - \mathbf{A}\mathbf{x}| \, |^2 + \\ &\sum_{i=1}^{M} \left(\ln \alpha_i - \alpha_i |x_i|^2 \right) + \sum_{i=1}^{M} \left((a-1) \ln \alpha_i + a \ln b - b\alpha_i \right) \\ &+ (c-1) \ln \gamma + c \ln d - d\gamma + \text{constants}. \end{aligned} \tag{7}$$

In the following, $c_{x_i}, c'_{x_i}, c_{\alpha_i}, c_\gamma, c_{a_{ji}}$ etc. represents normalization constants for the respective pdfs.

Update of $q_{x_i}(x_i)$: Using (6), $\ln q_{x_i}(x_i)$ turns out to be quadratic in x_i and thus can be represented as a Gaussian distribution as follows, Note that we split $\mathbf{A}\mathbf{x}$ as, $\mathbf{A}\mathbf{x} = \mathbf{C}_i x_i + \mathbf{C}_{\bar{i}} \mathbf{x}_{\bar{i}}$, where $\mathbf{C}_i$ represents the i^{th} column of $\mathbf{A}$, $\mathbf{C}_{\bar{i}}$ represents the matrix with i^{th} column of $\mathbf{A}$ removed, x_i is the i^{th} element of $\mathbf{x}$, and $\mathbf{x}_{\bar{i}}$ is the vector without x_i. In fact, we can represent $\mathbf{C}_i = (\bigotimes_{j=1}^{N} \mathbf{A}_{j,p_{ji}})$. To show the relation to the columns of the KS factor matrices $(p_1, p_2, ..., p_N)$ which generates $\mathbf{C}_i$, $i = 1 + \sum_{k=1}^{N} (p_k - 1) J_k$, $J_k = \prod_{m=N, m\neq i}^{k+1} P_m$, $P_{N+1} = 1$. So we denote $\mathbf{A}_{j,p_{ji}}$ as the column vector from $\mathbf{A}_j$ which generates $\mathbf{C}_i$. From the property of the Kronecker products [22] that $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$, we can verify that $\|\mathbf{C}_i\|^2 = (\bigotimes_{j=1}^{N} \mathbf{A}_{j,p_{ji}})^H (\bigotimes_{j=1}^{N} \mathbf{A}_{j,p_{ji}}) = \prod_{j=1}^{N} \left\|\mathbf{A}_{j,p_{ji}}\right\|^2$. Clearly, the mean and the variance of the resulting Gaussian distribution becomes,

$$\begin{aligned} \sigma_i^2 &= \frac{1}{<\gamma> \prod_{j=1}^{N} <\left\|\mathbf{A}_{j,p_{ji}}\right\|^2> + <\alpha_i>}, \ < x_i > = \\ \widehat{x}_i &= \sigma_i^2 ((< \mathbf{C}_i^H > \mathbf{y} - < (\mathbf{C}_i^H \mathbf{C}_{\bar{i}}) >< \mathbf{x}_{\bar{i}} >) < \gamma >, \end{aligned} \tag{8}$$

where $\widehat{x}_i$ represents the point estimate of x_i and $\widehat{\mathbf{A}}_{j,i} = [1 < \mathbf{a}_{j,i}^H >]^H$, $< \mathbf{a}_{j,i} >$ being the mean of $\mathbf{a}_{j,i}$ which follows from the below

derivation for $\mathbf{a}_{j,i}$. Also, note that in $< (\mathbf{C}_i^H \mathbf{C}_{\bar{i}}) >$, there are cross terms of the form $< \prod_{j=1}^{N} \mathbf{A}_{j,p_i}^H \mathbf{A}_{j,p_k} >, i \neq k$ which can be written as $= \prod_{j=1}^{N} < \mathbf{A}_{j,p_i}^H >< \mathbf{A}_{j,p_k} >$ because of the independence of the approximate distribution q of each columns of the factor matrices.

Update of $q_{\mathbf{a}_{j,i}}(\mathbf{a}_{j,i})$: Here we go back to the tensor representation. For simplicity, we define $\mathbf{V}_j =< \mathbf{X}^{(j)} >< (\overset{1}{\underset{k=N,k\neq j}{\bigotimes}} \mathbf{A}_k)^T >$, $\mathbf{W_j} =< \mathbf{X}^{(\mathbf{j})} (\overset{1}{\underset{\mathbf{k=N,k\neq j}}{\bigotimes}} \mathbf{A_k})^{\mathbf{T}} (\overset{1}{\underset{\mathbf{k=N,k\neq j}}{\bigotimes}} \mathbf{A_k})^* \mathbf{X}^{(\mathbf{j})\,\mathbf{H}} >$. The variational approximation for the vector $\mathbf{a}_{j,i}$ results in,

$$\ln q_{\mathbf{a}_{j,i}}(\mathbf{a}_{j,i}) = - < \gamma >< \left\{ \|\mathbf{Y} - [\![\mathbf{A}_1, ..., \mathbf{A}_N; \mathbf{x}]\!]\|^2 \right\} > \overset{(a)}{=} - < \gamma > \mathrm{tr}\Big\{ - \mathbf{Y}^{(j)} \mathbf{V}_j^H \mathbf{A}_j^H + \mathbf{A}_j \mathbf{V}_j \mathbf{Y}^{(j)} + \mathbf{A}_j \mathbf{W_j} \mathbf{A_j^H} \Big\} + \mathbf{c_{a_{ji}}} \tag{9}$$

In (a), we used the fact that [12] $\|\mathbf{A}\|^2 = \mathrm{tr}\{\mathbf{A}^{(k)} (\mathbf{A}^{(k)})^H\}$ for a tensor $\mathbf{A}$ and further we denote $\mathbf{A}_N \otimes ... \mathbf{A}_{j+1} \otimes \mathbf{A}_{j-1} ... \otimes \mathbf{A}_1 = \overset{1}{\underset{k=N,k\neq j}{\bigotimes}} \mathbf{A}_k$. In (9), $\mathrm{tr}\{\mathbf{A}_j \mathbf{W_j} \mathbf{A_j^H}\}$ can be written as, $\mathrm{tr}\{\mathbf{A}_{j,i} \mathbf{W_j} \mathbf{A_{j,i}^H}\}$+ "terms independent of $\mathbf{a}_{j,i}$", which gets simplified as $\mathrm{tr}\{\mathbf{W_j}\} \|\mathbf{a_{j,i}}\|^2$ + "others". Finally, the mean ($< \mathbf{a}_{ji} > = \widehat{\mathbf{a}}_{ji}$) and covariance ($\boldsymbol{\Upsilon}_{j,i}$) of the resulting Gaussian distribution can be written as (after expanding $\mathbf{V}_j, \mathbf{W_j}$),

$$\widehat{\mathbf{a}}_{ji} = (\mathbf{b}_j)_{\bar{1}}, \;\; \mathbf{b}_j = (\mathbf{Y}^{(j)} < \mathbf{X}^{(j)} >< (\overset{1}{\underset{k=N,k\neq j}{\bigotimes}} \mathbf{A}_k)^T >)_i,$$
$$\boldsymbol{\Upsilon}_{j,i} = \beta_{j,i} \mathbf{I}, \; \beta_{j,i} = \mathrm{tr}\{(\overset{1}{\underset{k=N,k\neq j}{\bigotimes}} < \mathbf{A}_k^T \mathbf{A}_k^* >) < \mathbf{X}^{(j)\,H} \mathbf{X}^{(j)} >\}, \tag{10}$$

where $(\cdot)_i$ represents the i^{th} column of the matrix $(\cdot)$ and $(\mathbf{b}_j)_{\bar{1}}$ represents the vector formed by all the elements except the first one of the vector $\mathbf{b}_j$. For the computation of the elements of the matrix $< \mathbf{X}^{(j)\,H} \mathbf{X}^{(j)} >$, the diagonal elements contain terms of the form $< |x_l|^2 >$ the expressions for which are provided below in (11). The non-diagonal terms contain terms of the form $< x_l x_k >, l \neq k$ which gets simplified due to the independence of the corresponding q distributions, $< x_l x_k > = \widehat{x}_l \widehat{x}_k$. Also, we can write $< \|\mathbf{A}_{j,i}\|^2 > = 1 + \|\widehat{\mathbf{a}}_{j,i}\|^2 + \beta_{j,i} I_j$, which gets used in (8).

Update of $q_{\alpha_i}(\alpha_i), q_\gamma(\gamma)$: The variational approximation leads to the Gamma distribution for the $q_{\alpha_i}(\alpha_i)$ and $q_\gamma(\gamma)$, which are parameterized by it's mean. The detailed derivation for this is omitted here, since it is provided in our paper [18]. The mean of the Gamma distribution for $q_{\alpha_i}(\alpha_i), q_\gamma(\gamma)$ is given by,

$$< \alpha_i > = \frac{a + \frac{1}{2}}{\left(< |x_i|^2 > + b \right)}, \; < \gamma > = \frac{c + \frac{N}{2}}{\left(< \left\| \mathbf{y} - (\overset{N}{\underset{j=1}{\bigotimes}} \mathbf{A}_j) \mathbf{x} \right\|^2 > + d \right)}, \tag{11}$$

where, $< \|\mathbf{y} - (\overset{N}{\underset{j=1}{\bigotimes}} \mathbf{A}_j)\mathbf{x}\|^2 > = \|\mathbf{y}\|^2 - 2\mathbf{y}^H (\overset{N}{\underset{j=1}{\bigotimes}} < \widehat{\mathbf{A}}_j >) \widehat{\mathbf{x}} + \mathrm{tr}((\overset{N}{\underset{j=1}{\bigotimes}} < \mathbf{A}_j^H \mathbf{A}_j >)(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^H + \boldsymbol{\Sigma})), \boldsymbol{\Sigma} = \mathrm{diag}\,(\sigma_1^2, ..., \sigma_M^2)$, $\widehat{\mathbf{x}} = [\widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_M]^H$. and $< |x_i|^2 > = |\widehat{x}_i|^2 + \sigma_i^2$,. From (8), it can be seen that the estimate $\widehat{\mathbf{x}}$ converges to the L-MMSE equalizer, $\widehat{\mathbf{x}} = (\mathbf{A}^H \mathbf{A} + \frac{1}{<\gamma>} \boldsymbol{\Sigma}^{-1})^{-1} \mathbf{A}^H \mathbf{y}$. This version of the SAVE where each columns of the factor matrices are updated independently is called as SAVED-KS (SAVE with KS Dictionary learning).

3.1. Joint VB for KS Dictionary Learning

In this section, we treat the columns of the factor matrix $\mathbf{A}_j$ jointly in the approximate posterior using VB. We also define for the convenience of the analysis, $\mathbf{A}_j = [\mathbf{1}\, \mathbf{A}_{\bar{1},j}^H]^H$, where $\mathbf{A}_{\bar{1},j}$ represents all other rows except the first and $\mathbf{1}$ represents a column vector (of size P_j) with all ones. $\ln q_{\mathbf{A}_j}(\mathbf{A}_j) = \mathrm{tr}\{-\mathbf{Y}^{(j)} \mathbf{V}_j^H \mathbf{A}_j^H - \mathbf{A}_j \mathbf{V}_j \mathbf{Y}^{(j)\,H} + \mathbf{A}_j \mathbf{W_j} \mathbf{A_j^H}\} + \mathbf{c_{A_j}}$, Defining $\mathbf{B}_j$ as with the first row of $(\mathbf{Y}^{(j)} \mathbf{V}_j^H)$ removed. So $\mathrm{tr}\{-\mathbf{Y}^{(j)} \mathbf{V}_j^H \mathbf{A}_j^H\} = \sum_{i=1}^{M} (\mathbf{Y}^{(j)} \mathbf{V}_j^H)_{1,i} + \mathrm{tr}\{\mathbf{B}_j \mathbf{A}_{\bar{1},j}^H\}$, $(\mathbf{Y}^{(j)} \mathbf{V}_j^H)_{1,i}$ represents the $(1,i)^{th}$ element of the matrix. Now expanding the term $\mathbf{A}_j \mathbf{W_j} \mathbf{A_j^H} = [\mathbf{1}\, \mathbf{A_{\bar{1},j}^H}]^{\mathbf{H}} \mathbf{W_j} [\mathbf{1}\, \mathbf{A_{\bar{1},j}^H}]$ which simplifies $\ln q_{\mathbf{A}_j(\mathbf{A}_j)}$ as, $\ln q_{\mathbf{A}_j}(\mathbf{A}_j) = < \gamma > \mathrm{tr}\{\mathbf{B}_j \mathbf{A}_{\bar{1},j}^H\} + < \gamma > \mathrm{tr}\{\mathbf{A}_{\bar{1},j} \mathbf{B}_j^H\} - < \gamma > \mathrm{tr}\{\mathbf{A}_{\bar{1},j} \mathbf{W_j} \mathbf{A_{\bar{1},j}^H}\}$. This corresponds to the functional form of a circularly-symmetric complex matrix normal distribution [27]. This can be represented for a random matrix $\mathbf{X} \in \mathbf{C}^{n \times p}$ as $p(\mathbf{X}) \propto \exp(-\mathrm{tr}\{\boldsymbol{\Psi}^{-1}(\mathbf{X} - \mathbf{M})^H \boldsymbol{\phi}^{-1} (\mathbf{X} - \mathbf{M})\})$, which is denoted as $\mathcal{CMN}(\mathbf{X} \mid \mathbf{M}, \boldsymbol{\phi}, \boldsymbol{\Psi})$. Thus the variational approximation for $\mathbf{A}_{\bar{1},j}$ gets represented as $\mathcal{CMN}(\mathbf{A}_{\bar{1},j} \mid \mathbf{M}_j, \mathbf{I}_M, \boldsymbol{\Psi}_j)$.

$$\mathbf{M}_j = \widehat{\mathbf{A}}_{\bar{1},j} = < \gamma > \mathbf{B}_j \boldsymbol{\Psi}_j, \; \boldsymbol{\Psi}_j = (< \gamma > \mathbf{W_j})^{-\mathbf{1}} \tag{12}$$

Note that $vec(\mathbf{A}_{\bar{1},j}) \sim \mathcal{N}(vec(\mathbf{M}_j), \boldsymbol{\Psi}_j \otimes \mathbf{I}_M)$, so the terms of the form $< \|\mathbf{A}_{j,i}\|^2 >$ in (8) becomes, $< \|\mathbf{A}_{j,i}\|^2 > = 1 + \|\mathbf{M}_{j,i}\|^2 + (\boldsymbol{\Psi}_j)_{i,i}$. $(\boldsymbol{\Psi}_j)_{i,i}$ is the i^{th} diagonal element of $\boldsymbol{\Psi}_j$ and $\mathbf{M}_{j,i}$ represents the i^{th} column of $\mathbf{M}_j$. Also, we can represent $\mathbf{A}_j^H \mathbf{A}_j = \mathbf{1}\mathbf{1}^H + \mathbf{M}_j^H \mathbf{M}_j + (I_j - 1)\boldsymbol{\Psi}_j$.

For our proposed SAVED-KS, it is evident that we do not need any matrix inversions compared to [28,29]. Update of all the variable $\mathbf{x}, \boldsymbol{\alpha}, \gamma$ involves simple addition and multiplication operations. We also introduce the following notations, $\mathbf{x}_{i-} = [x_1 ... x_{i-1}]^T, \mathbf{x}_{i+} = [x_{i+1} ... x_M]^T$.

Algorithm 1 SAVED-KS SBL Algorithm

Given: $\mathbf{y}, \mathbf{A}, I_j, P_j \; \forall j$.

Initialization: a, b, c, d are taken to be very low, on the order of 10^{-10}, thus $p(\alpha_i) \propto \alpha_i^{-1}, p(\gamma) \propto \gamma^{-1}$ which corresponds to a non-informative Jeffrey's prior [30]. $\alpha_i^0 = a/b, \forall i, \gamma^0 = c/d$ and $\sigma_i^{2,0} = \frac{1}{\|\mathbf{C}_i^0\|^2 \gamma^0 + \alpha_i^0}, \mathbf{x}^0 = \mathbf{0}$. Random initialization for the dictionary matrix $\mathbf{A}_j \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$.

At iteration $t+1$ (superscript t is used to denote the iteration stage),

1. Update $\sigma_i^{2,t+1}, \widehat{x}_i^{t+1}, \forall i$ from (8) using $\mathbf{x}_{i-}^{t+1}$ and $\mathbf{x}_{i+}^t$.
2. Update $\widehat{\mathbf{A}}_{j,i}^{t+1}, \boldsymbol{\Upsilon}_{j,i} \; \forall i, j$ from (10) or $\widehat{\mathbf{A}}_j^t, \boldsymbol{\Psi}_j$ from (12).
3. Compute $< x_i^{2,t+1} >$ from (11) and update α_i^t, γ^{t+1}.
4. Continue steps $1-4$ till convergence of the algorithm.

4. IDENTIFIABILITY OF KS DICTIONARY LEARNING

The local identifiability (upto permutation ambiguity) of the KS DL is ensured if the FIM is non-singular [31]. We can write $\mathbf{A}_{j,i} = \mathbf{F}_j^{(i)} \boldsymbol{\theta}_j, \boldsymbol{\theta}_j = vec(\mathbf{A}_j)$ and $\mathbf{F}_j^{(i)} = [\mathbf{0}_{I_j \times I_j (i-1)} \; \mathbf{I}_{I_j} \; \mathbf{0}_{I_j \times I_j (P_j - i)}]$ and we define $\mathbf{F}_r = \overset{}{\underset{p_{ji}, \forall j}{\bigotimes}} \mathbf{F}_j^{(p_{ji})}, r = \sum_{j=1}^{N} (p_{ji} - 1) J_j + p_{Ni}, J_j = \prod_{r=j+1}^{N} P_r$. We observe that we can separate the contributions of $\boldsymbol{\theta}$ and $\mathbf{x}$ in (3) as, $\mathbf{y} = \underbrace{(\sum_{r=1}^{M} x_r \mathbf{F}_r)}_{\mathbf{F}(\mathbf{x})} \underbrace{(\bigotimes_{j=1}^{N} \boldsymbol{\theta}_j)}_{\mathbf{f}(\boldsymbol{\theta})} + \mathbf{w}$. Writing

$\ln p(\mathbf{y},\boldsymbol{\theta},\mathbf{x}) = \ln p(\mathbf{y}/\mathbf{x},\boldsymbol{\alpha},\gamma,\mathbf{A}) + \ln p(\mathbf{x}/\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\gamma)$, it is clear that the FIM can be split as $FIM = FIM_y + FIM_{prior}$. For FIM_y, extending the derivation of the CRB for the KS dictionary matrices in [31] to the high-order tensor SBL case, we define the Jacobian matrix of $\mathbf{S} = \mathbf{F}(\mathbf{x})\mathbf{f}(\boldsymbol{\theta})$ as,

$$\begin{aligned}&\mathbf{J}(\boldsymbol{\theta},\mathbf{x}) = [\mathbf{J}(\boldsymbol{\theta})\ \mathbf{J}(\mathbf{x})],\ \mathbf{J}(\boldsymbol{\theta}) = [\mathbf{J}(\boldsymbol{\theta_1}) ..... \mathbf{J}(\boldsymbol{\theta_N})] \text{ where,}\\ &\mathbf{J}(\boldsymbol{\theta_j}) = \mathbf{F}(\mathbf{x})(\boldsymbol{\theta_1} \otimes ... \mathbf{I_{I_j P_j}} .... \otimes \boldsymbol{\theta_N}),\\ &\mathbf{J}(\mathbf{x}) = [\mathbf{F_1}(\bigotimes_{\mathbf{j=1}}^{\mathbf{N}} \boldsymbol{\theta_j}),\ ....,\ \mathbf{F_M}(\bigotimes_{\mathbf{j=1}}^{\mathbf{N}} \boldsymbol{\theta_j}))].\end{aligned} \tag{13}$$

We define $\boldsymbol{\Gamma} = \text{diag}\,(\boldsymbol{\alpha})$. Further, the FIM for the case of SBL can be written as,

$$FIM = \begin{bmatrix} \mathrm{E}(\gamma)\mathbf{J}(\boldsymbol{\theta})^{\mathbf{H}}\mathbf{J}(\boldsymbol{\theta}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{E}(\gamma)\mathbf{J}(\mathbf{x})^{\mathbf{H}}\mathbf{J}(\mathbf{x}) + \mathrm{E}(\boldsymbol{\Gamma}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & a\,\mathrm{E}(\boldsymbol{\Gamma}^{-2}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (N+c-1)\,\mathrm{E}(\gamma^{-2}) \end{bmatrix} \tag{14}$$

Here, $\gamma\mathbf{J}(\mathbf{x})^{\mathbf{T}}\mathbf{J}(\boldsymbol{\theta}) = \mathbf{0}$, since $\mathbf{x}$ is zero mean. Further using the expression for the inverse of the block FIM above, for non-singularity, $\mathbf{J}(\boldsymbol{\theta})$ should be full rank. For the FIM analysis, we assume that the support (no. of non-zero elements of $\mathbf{x}$) is known, then $\mathrm{E}(\gamma)\mathbf{J}(\mathbf{x})^{\mathbf{H}}\mathbf{J}(\mathbf{x}) + \mathrm{E}(\boldsymbol{\Gamma})$ and $a\,\mathrm{E}(\boldsymbol{\Gamma}^{-2})$ becomes invertible if $\prod_{j=1}^{N} I_j > K$. Assuming $\prod_{j=1}^{N} I_j > \sum_{j=1}^{N}(I_j - 1)P_j$ ($I_j - 1$ since the columns are scaled to make the first entry 1), i.e. no. of degrees of freedom in the dictionary $< \prod_{j=1}^{N} I_j$, then it is clear FIM is non-singular. Another remark is that here we consider only single measurement vector case and it is evident from the FIM expression that it can be non-singular even in this case under certain conditions on the dimensions of the KS factor matrices. We also observe that identifiability results for a mix of structured (Vandermonde matrices) and unstructured KS matrices for $3-$way tensors are discussed in [32]. Note that algorithms which deal with KS dictionary matrices are very recent and fundamental limits of the estimation accuracy for such systems in a minimax setting can be seen in [33].

4.1. Identifiability for mix of parametric and non-parametric KS factors

We briefly outline the results for the case of mixture of parametric and non-parametric KS factors. We assume that the parameters $\mathbf{A}_j, j = 1, ..., P, P < N$ are Vandermonde matrices parameterized by the spatial response $\phi_{j,l}, l = 1, ..., P_j$ and $\mathbf{A}_{j,l} = [1\, e^{ig_j(\phi_{j,l})} .... e^{i(I_j-1)g_j(\phi_{j,l})}]^T, i = \sqrt{-1}$, where for e.g. $g_j(\phi_{j,l}) = \pi \sin(\phi_{j,l})$ and angles are sufficiently separated such that each of the columns $\mathbf{A}_{j,l}$ becomes linearly independent. This corresponds to the case of antenna array response for ULA or frequency response parameterized by a delay. Further vectorizing $\boldsymbol{\theta}_j = vec(\phi_{j,1}, ..., \phi_{j,P_j})$, so the degrees of freedom reduces to P_j instead of $I_j P_j$ for the unstructured case. So, $\forall j = 1, ..., P$,

$$\mathbf{J}(\boldsymbol{\theta_j}) = \mathbf{F_{pa}}(\mathbf{x})(\boldsymbol{\theta_1} \otimes \ ... \mathbf{E_j}\mathbf{A_j}\mathbf{F_j} \ ... \boldsymbol{\theta_P} \otimes \boldsymbol{\theta_{P+1}} .... \otimes \boldsymbol{\theta_N}), \tag{15}$$

where $\mathbf{F}_{pa}(\mathbf{x})$ has the same expression as $\mathbf{F}(\mathbf{x})$ with $\mathbf{F}_j^{(i)}, \forall j = 1, .., P$ becomes a matrix with all ones of size $I_j \times I_j$, $\mathbf{E_j} = \text{diag}\,(\mathbf{0}, \mathbf{1}, ..., (\mathbf{I_j} - \mathbf{1}))$ and $\mathbf{F}_j = i\,\text{diag}\,(g_j'(\phi_{j,1}), ..., g_j'(\phi_{j,P_j}))$. Thus for parametric factors, $\mathbf{J}(\boldsymbol{\theta_j})$ becomes a vector of size $\prod_j I_j \times P_j$. The identifiability conditions can be restated as, assuming $\prod_{j=1}^{N} I_j > \sum_{j=1}^{P} P_j + \sum_{j=P+1}^{N}(I_j - 1)P_j$, i.e. no. of degrees of freedom in the dictionary $< \prod_{j=1}^{N} I_j$, then it is clear FIM is non-singular.

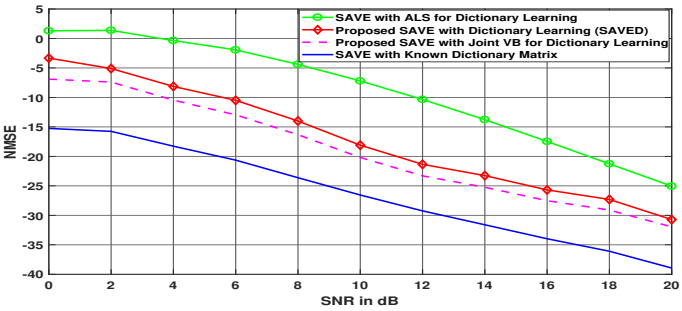

Fig. 1. NMSE vs SNR in dB.

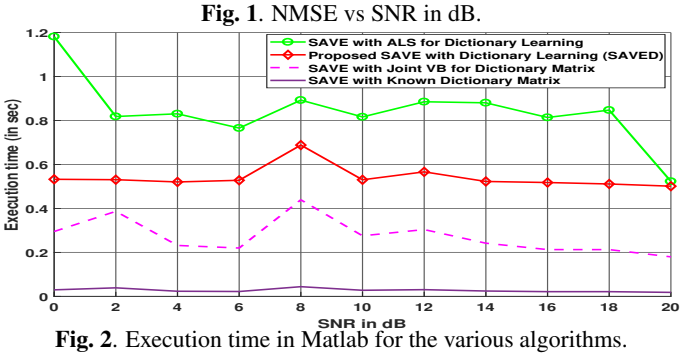

Fig. 2. Execution time in Matlab for the various algorithms.

5. SIMULATION RESULTS

In this section, we present the simulation results to validate the performance of our SAVED-KS SBL algorithm (Algorithm 1) compared to state of the art solutions. For the simulations, we consider a $3-D$ tensor with dimensions $(4, 4, 4)$ and the number of non-zero elements of $\mathbf{x}$ or the rank of the tensor (no of non-zero elements of $\mathbf{x}$) is fixed to be 4. All the elements of the dictionary matrix $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ and non-zero elements of $\mathbf{x}$ are generated i.i.d complex Gaussian, $\mathcal{CN}(0, 1)$ and the singular values are modified to convert the matrices such that they have a particular condition number $(= 2)$. This is done to ensure that the system identifiability is not affected by the Krushkal ill-conditioning [12]. Normalized Mean Square Error (NMSE) is defined as $NMSE = \frac{1}{M}||\widehat{\mathbf{x}} - \mathbf{x}||^2$, $\widehat{\mathbf{x}}$ represents the estimated value, $NMSE_{dB} = 10\log 10(NMSE)$. In Figure 1, we depict the normalized MSE (NMSE) performance of our proposed SAVED-KS algorithm with the classical ALS algorithm which doesn't utilize any statistical information about the dictionary or sparse coefficients. Our SAVED-KS algorithm has much better reconstruction error performance compared to the ALS [12] and our joint VB version performs better than the SAVED-KS version, but comes with a higher computational complexity due to the matrix inversion. It is clear from Figure 2 that proposed SAVE approach has a faster convergence rate than the ALS.

6. CONCLUSION

We presented a fast SBL algorithm called SAVED-KS, which uses the variational inference techniques to approximate the posteriors of the data, hyper-parameters and the factor matrices of the dictionary. We showed that the proposed algorithm has a faster convergence rate and better performance in terms of NMSE than even the state of the art ALS solutions for dictionary learning. Possible extensions to the current work might include: i) Convex combination of structured and unstructured KS factor matrices, for e.g., DoA response closeness to the vandermonde. ii) Asymptotic performance analysis and mismatched Cramer-Rao bounds [34] for the SAVED-KS algorithm.

7. REFERENCES

[1] D. Nion and N. D. Sidiropoulos, "Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar ," *IEEE Trans. on Sig. Process.*, vol. 58, no. 11, Nov. 2010.

[2] C. Qian, X. Fu, N. D. Sidiropoulos, and Y. Yang, "Tensor-based parameter estimation of double directional massive MIMO channel with dual-polarized antennas," in *ICASSP*, 2018.

[3] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis," in *UCLA Working Papers in Phonetics, Available at http://publish.uwo.ca/harshman/wpppfac0.pdf.*, 1970.

[4] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change ," *Probl. Meas. Change*, 1963.

[5] M. Srensen, L. D. Lathauwer, P. Comon, S. Icart, and L. Deneire, "Canonical polyadic decomposition with a columnwise orthonormal factor matrix ," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 4, 2010.

[6] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing ," *Proceedings of the IEEE*, vol. 21, no. 2, Feb. 2012.

[7] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit ," *IEEE Trans. Info. Theory*, vol. 53, no. 12, December 2007.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit ," *SIAM J. Sci. Comput.*, vol. 20, no. 1, 1998.

[9] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions ," *IEEE J. Sel. Top. Sig. Process.*, vol. 4, no. 2, April 2010.

[10] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, 2001.

[11] D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Trans. on Sig. Process.*, vol. 52, no. 8, August 2004.

[12] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications ," *SIAM Review*, vol. 51, no. 2, Aug. 2009.

[13] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete respresentations ," *Neural Computation*, vol. 12, no. 2, 2000.

[14] K. Skretting and K. Engang, "Recursive least squares dictionary learning algorithm ," *IEEE Trans. on Sig. Process.*, vol. 58, Apr. 2010.

[15] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation ," *IEEE Trans. on Sig. Process.*, vol. 54, no. 11, Nov. 2006.

[16] F. Roemer, G. D. Galdo, and M. Haardt, "Tensor-based algorithms for learning multidimensional separable dictionaries," in *IEEE Intl. Conf. on Acous., Speech and Sig. process. (ICASSP)*, 2014.

[17] X. Ding, W. Chen, and I. J. Wassell, "Joint sensing matrix and sparsifying dictionary optimization for tensor compressive sensing. ," *IEEE Trans. on Sig. Process.*, vol. 65, no. 4, Jul. 2017.

[18] C. K. Thomas and D. Slock, "SAVE - Space alternating variational estimation for sparse Bayesian learning," in *IEEE Data Science Workshop*, June 2018.

[19] ——, "Space Alternating Variational Bayesian Learning for LMMSE Filtering," in *IEEE Eur. Sig. Process. Conf. (EUSIPCO)*, September 2018.

[20] ——, "Gaussian variational Bayes Kalman filtering for dynamic sparse Bayesian learning," in *IEEE 5th Intl. Conf. on Time Ser. and Forecast. (ITISE)*, September 2018.

[21] ——, "SAVED - Space alternating variational estimation for sparse Bayesian learning with parametric dictionaries," in *52nd IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Oct. 2018.

[22] N. D. Sidiropoulos *et al.*, "Tensor decomposition for signal processing and machine learning ," *IEEE Trans. on Sig. Process.*, vol. 65, no. 13, July 2017.

[23] Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Identifiability of Kronecker-structured dictionaries for tensor data ," *IEEE Journ. Sel. Top. in Sig. Process.*, vol. 12, no. 5, Oct. 2018.

[24] C. K. Thomas and D. Slock, "Variational Bayesian learning for channel estimation and transceiver determination," in *IEEE Info. Theo. and Appl. Wkshp. (ITA)*, Feb 2018.

[25] M. J. Beal, "Variational algorithms for approximate bayesian inference," in *Thesis, Univeristy of Cambridge, UK*, May 2003.

[26] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Sig. Process. Mag.*, vol. 29, no. 6, pp. 131–146, November 2008.

[27] A. K. Gupta and D. K. Nagar, "Matrix variate distributions," in *Boca Raton FL, USA: CRC Press*, 1999.

[28] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast variational sparse bayesian learningwith automatic relevance determination for superimposed signals," *IEEE Trans. on Sig. Process.*, vol. 59, no. 12, December 2011.

[29] M. Al-Shoukairi, P. Schniter, and B. D. Rao, "GAMP-based low complexity sparse bayesian learning algorithm," *IEEE Trans. on Sig. Process.*, vol. 66, no. 2, January 2018.

[30] K. P. Murphy, "Machine learning: A probabilistic perspective," in *MA, USA: MIT press*, 2012.

[31] M. Boizard, R. Boyer, G. Favier, and P. Comon, "Performance estimation for tensor CP decomposition with structured factors," in *IEEE Intl. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Brisbane, Australia, 2015.

[32] M. Sorensen and L. D. Lathauwer, "Blind Signal Separation via Tensor Decomposition With Vandermonde Factor: Canonical Polyadic Decomposition ," *IEEE Trans. on Sig. Process.*, vol. 61, no. 22, 2013.

[33] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds on dictionary learning for tensor data ," *IEEE Trans. Info. Theory*, vol. 64, no. 4, Apr. 2018.

[34] C. D. Richmond and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification ," *IEEE Trans. on Sig. Process.*, vol. 63, no. 9, May. 2015.